Chapter 9. Models

Fitting models to data is a popular activity. For data taking values in a group or homogeneous space, the associated representation theory gives neat families of models. Briefly, if $P(x)$ is a probability on X, write $P(x) = e^{h(x)}$ with $h = \log P$. Then expand h in a natural basis b_i of $L(X)$: $h(x) = \Sigma\theta_i b_i(x)$. Any positive probability can be expanded in this way. Truncating the expansion to a fixed number of terms leads to families of simple measures or models; because b_i are orthogonal, θ_i are identifiable.

It is an interesting fact that many models introduced by applied workers fall into this class. The general story is presented first, then a specialization to data on spheres, then a specialization to partially ranked data. A brief review of other approaches to ranked data is followed by a section supplying the relevant exponential family theory.

A. Exponential families from representations

Let G be a group acting transitively on a compact set X. Let $L(X)$ denote the real valued continuous functions on X. Suppose X has an invariant distribution dx. The following abstracts an idea introduced by Lo (1977) and Beran (1979).

Definition. Let Θ be an invariant subspace of $L(X)$ containing the constants. Define a family of measures, one for each $\theta \in \Theta$, by specifying the densities to be

$$P_\theta(dx) = a(\theta)e^{\theta(x)}dx,$$

where $a(\theta)$ is a normalizing constant forcing $P_\theta(dx)$ to integrate to 1.

Suppose Θ is finite dimensional. Let b_0 =constant, b_1, $b_2, \ldots, b_p$ be a basis for Θ. Then the family can be parameterized as

$$(*) \qquad P_\theta(dx) = a(\theta)e^{\theta' b}dx, \ \theta \in \mathbb{R}^p, b = (b_1(x), \ldots, b_p(x)).$$

Lemma 1. *The family * is well parameterized in the sense that $P_\theta = P_{\theta'}$ if and only if $\theta = \theta'$.*

Proof. Only the forward direction requires proof. If $P_\theta = P_{\theta'}$, then

$$(\theta - \theta') \cdot b(x) = \log(a(\theta')/a(\theta)) \text{ for all } x.$$

The left side is a linear combination of b_1, $b_2, \ldots, b_p$ which is constant. But $1, b_1$, $b_2, \ldots, b_p$ is a basis, so $\theta = \theta'$. □

In applications there is a decomposition into invariant subspaces $L(X) = V_0 \oplus V_1 \oplus V_2 \oplus \ldots$ and Θ's are chosen as a finite direct sum of subspaces. Usually

these nest together neatly to form zeroeth order models (the uniform distribution), 1st order models, etc. The matrix entries of the irreducible representations then provide a convenient basis.

The easiest example is for data on Z_2^k. The exponential families that the group theory suggests are exactly the log-linear models that statisticians fit to $2 \times 2 \ldots \times 2$ tables (k factors). Here a person is classified via k dichotomous variables. This gives rise to a vector in Z_2^k or locates a cell in the table.

A useful entry to the statistical literature is provided by the first few chapters of Gokhole and Kullback (1978). General contingency tables can be treated similarly. Since this is such a well studied area, we will not pursue it further than mentioning the important paper of Darroch, Lauritzen and Speed (1980). This gives an elegant interpretation to setting $\theta_i = 0$ for a large class of models. It would be an important contribution to generalize their ideas to the general group case.

The measure $P_\theta(dx)$ governs a single observation. We model a sample of size n by a product measure

$$P_\theta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(x_i) dx_i.$$

The statistical problem becomes, given a model Θ, and observations x_1, $x_2, \ldots, x_n$, what is a reasonable guess for θ, and how sure are you about the answer.

Remark 1. Crain (1973, 1974, 1976) suggested expansion of log $P(x)$ in a basis of orthogonal functions as a route to nonparametric density estimation. He truncated the expansion at a point depending on sample size. This leads to an approximate density in a finite dimensional exponential family as in the definition above.

Crain's later papers give conditions on how large the cutoff point should be to have the maximum likelihood estimator exist. These are discussed in Section E below.

Remark 2. There is a growing literature on orthogonal series estimators – density estimators based on expanding the density directly as $P(x) = \Sigma \theta_i b_i(x)$. Hall (1986) makes noteworthy contributions providing simple useable estimators and giving sharp rates of convergence. He gives pointers to the literature. Hall's results can be carried over to problems on compact homogeneous spaces in a straightforward way.

Orthogonal series estimators suffer from the possibility of negative density estimates. This is why Crain worked with log $P(x)$. It is a worthwhile project to combine the ideas and bounds of Hall with the ideas of Crain.

Remark 3. One problem encountered with log P: it is badly behaved if $P = 0$. Consider a density on the circle. If $P(x) > 0$ outside an interval, things can be rescaled and there is no trouble. If $P(x) = 0$ on several intervals the problem can be treated as a mixture, but the log P approach is wearing out its welcome.

There are so many other density estimates possible – from histograms, through kernel estimators, through projection pursuit.

On more general homogeneous spaces, problems with vanishing density seem even less approachable.

Remark 4. The definition above is in terms of real valued functions. This works fine for the symmetric group and its homogeneous spaces and for the orthogonal group. In general, $L(X)$ may be taken as all complex functions and a model may be taken as an invariant subspace of $L(X)$. Just as any real function on Z_n can be expanded in terms of $\sin(2\pi jk/n)$ and $\cos(2\pi jk/n)$, any real function on X can be expanded as a real linear combination of the real and imaginary parts of the matrix entries of the irreducible representations that occur in the splitting of $L(X)$.

Remark 5. The models introduced here blend in nicely with the spectral theory of Chapter 8. They are the largest models which allow as sufficient statistics the ingredients of the matching spectral analysis. See E-1 below.

Remark 6. The ideas set out above can be generalized in various ways. One natural extension begins with a space X and a symmetric Markov chain $P(x, dx)$ on X. Symmetric chains can be orthogonally diagonalized, and the eigen vectors provide a convenient orthogonal basis for $L(X)$. There are chains that don't arise from groups where this basis can be written explicitly. See Banni and Ito (1986, 1987) or Diaconis and Smith (1987). It is not clear if these models can be connected to the underlying chain.

A word of caution: I find the statistical community introduces models much too easily. In some cases, there is a justification: "height is the sum of a lot of small factors, so heights should be approximately normally distributed" or "the number of accidents is the sum of a lot of roughly independent binomial variables with small parameters, so accidents should be approximately Poisson." In some cases linearity or physical justification (and repeated comparison with reality) justify models: Gauss' discovery of Ceres, Bright-Wigner distributions in particle physics or multinomial distributions in genetics are examples.

The cases where some slim justification is given seem alarmingly few to me. Usually, one is contemplating some data and a model is chosen for convenience as a way of doing data analysis. This is a curve fitting approach and is fine, except that the product model assumes independence. Further, the assumptions about $P_\theta(dx)$ may be a drastic oversimplification. One may well do better looking directly at the data using spectral analysis, or a convenient ad hoc approach.

I must admit that I too find ad hoc modeling attractive and occasionally useful – it seems like a most worthwhile project to try to isolate what good comes out of the modeling paradigm and attempt to build a theory that optimizes this good instead of behavior in a non-existent fantasy land of iid repetitions.

B. Data on spheres.

Spherical data is discussed in Chapter 5-C. One important special problem is testing for uniformity. A large number of special tests have been suggested. These are reviewed by Mardia (1972) and Watson (1983). We discuss here one of the earliest tests and follow its later developments.

Let $X_1, X_2, \ldots, X_n$ be unit vectors on the sphere S^p in p dimensions. Define the sample resultant $\overline{R}$ and sample mean direction $U(\overline{\theta})$ by

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{R}U(\overline{\theta}).$$

Intuitively, if X_i are uniform, $\overline{R}$ will be "small" because there will be a lot of cancellation. If X_i are non-uniform and cluster around some point, then $\overline{R}$ will be "large." Rayleigh (1919) worked out the distribution of $\overline{R}$ under the uniform distribution and could thus propose a test "reject uniformity if $\overline{R} > r$" where r is chosen to achieve a given proportion of false rejections. A nice derivation of Rayleigh's results is given by Feller (1971, pg. 32).

Questions and alternate tests immediately suggest themselves. Observe that Rayleigh's test is invariant: $\overline{R}$ does not change if $X_1, X_2, \ldots, X_n$, are replaced by $\Gamma X_1, \ldots, \Gamma X_n$, Γ orthogonal. On the negative side, Rayleigh's test would not be appropriate if the X_i tend to cluster either close to a point *or* its antipode. When is this test a good one? Some answers have come from statistical theory.

Independent of Rayleigh, a class of natural non-uniform distributions was developed and used by Von Mises and Fisher. These have the form

$$P_{\mu,k}(dx) = C_p(k)e^{k\mu' x}dx$$

with $x \in S^p$, dx the uniform distribution, $\mu \in S^p$, and $k \geq 0$. The normalizing constant is

$$C_p(k) = k^{(p-1)/2}/(2\pi)^{p/2}I_{(p-1)/2}(k)$$

with $I_r(k)$ the modified Bessel function of the first kind.

The $P_{\mu,k}$ have "mean direction" μ and as k increases are more and more concentrated about μ. They arise naturally from the first hitting place of a Brownian particle with drift on the sphere. Watson (1983, Chapter 3) discusses this and other justifications.

A nice result is that the likelihood ratio test of

$$H_0: k = 0 \text{ vs. } H_1: k > 0, \ \mu \text{ unknown}$$

reduces to Rayleigh's test. Further, Rayleigh's test is the uniformly most powerful invariant test of uniformity versus $P_{\mu,k}(dx)$. These results are due to Beran (1968) who discusses their analog on compact homogeneous spaces. Giné (1975), Wellner (1979), and Jupp and Spurr (1985) amplify and develop these ideas. Closely related developments in the signal processing literature are surveyed by Lo and Eshleman (1979).

These developments give a pleasing answer to the original question: when is Rayleigh's test good – it's good if data cluster about one point μ in a spherically symmetric way.

Remark. I cannot resist reporting some background on Fisher's motivation for working with the distribution discussed above. This story was told to me in 1984 by the geologist Colin B. B. Bull. Dr. Bull was a student in Cambridge in the early 1950's. One day he ran across the street in haste and knocked an old man off a bicycle! The old man seemed dazed. When asked where he was bound he replied "India." It turned out to be R. A. Fisher who was meeting a train enroute to a visit to the Indian Statistical Institute. A month later, Bull met Fisher at Cambridge and again apologized. Fisher asked what area Bull worked in. Bull explained that a group of geologists was trying to test Wegener's theory of continental drift. Wegener had postulated that our current continents used to nest together. He tested this by looking at the distribution of a wide variety of bird, animal and plant life – arguing that matching points had close distributions.

Geologists found themselves far afield in trying to really understand Wegener's arguments. They searched for data that were closer to geology. They had hit on the distribution of magnetization angle in rocks. This gave points naturally distributed on the sphere. They had two distributions (from matching points on two continents) and wanted to test if the distributions were the same.

Fisher took a surprisingly keen interest in the problem and set out to learn the relevant geology. In addition to writing his famous paper (which showed the distributions were different) he gave a series of talks at the geology department to make sure he'd got it right. Bull told me these were very clear, and remarkable for the depth Fisher showed after a few months study.

Why did Fisher take such a keen interest? A large part of the answer may lie in Fisher's ongoing war with Harold Jeffries. They had been rudely battling for at least 30 years over the foundations of statistics. Jeffries has never really accepted (as of 1987!) continental drift. It is scarcely mentioned in Jeffries' book on geophysics. Fisher presumably had some extra-curricular motivation.

The motivation for Rayleigh's and Von Mises' work seems equally fascinating! Watson (1983, Chapter 3) gives a good set of pointers.

There is a second family of probabilities on S^p that has received a good deal of attention. The *Bingham densities* are defined on S^p as

$$b_p(D)\ \exp\{tr[DR'xx'R]\}dx$$

where D is a $p \times p$ diagonal matrix with (p,p) entry zero, and R is a $p \times p$ orthogonal matrix.

These densities are invariant under $x \to -x$ and so are possible models for unsigned directional data – lines in $\mathbb{R}^p$ (or points in projective space). A host of properties and characterizations of these densities are known.

Beran (1979) points out that both the Fisher-Von Mises and Bingham families fit nicely with the definition of models given in Section A. Here, the group $SO(p)$ of $p \times p$ orthogonal matrices with determinant 1 operates transitively on the space $X = S^p$. Take $L(X)$ as the continuous real valued functions on X.

Let P_k be the homogeneous polynomials (in $\mathbb{R}^p$) of degree k. Let M_k be the subspace of harmonic functions in P_k: $M_k = \{f: \nabla^2 f = 0\}$ where $\nabla^2 = \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2}$. These M_k are invariant and irreducible under the action of SO_p. Further, $L(X) = \oplus_{k=0} M_k$ as a Hilbert space direct sum. Proofs are in Dunkl and Ramirez (1971).

Following the definition, M_0 – the zero-th order model gives only the uniform distribution. $M_0 \oplus M_1$ – the first order models is obviously spanned by $1, x_1, x_2, \ldots, x_p$ (these are all killed by ∇^2). The associated exponential family is the Fisher-Von Mises family.

A second-order model is defined by $M_0 \oplus M_1 \oplus M_2$. Beran (1979) shows these are spanned by $\{x_i x_j\} - \{x_p^2\}$, giving the Bingham distribution. In general, a basis for $\bigoplus_{k=0}^{r} M_k$ consists of all distinct monomials of degree r and $r-1$, excluding x_p^r if r is even or x_p^{r-1} if r is odd.

Some more technical discussion of estimates and their properties is given in Section E below.

C. Models for permutations and partially ranked data.

Begin with a data set on the symmetric group S_n. Say $f(\pi)$ is the proportion of the data choosing rankng π. In working with such data it seems natural to begin by looking at first order statistics: the proportions ranking each item first, or last, and more generally the proportion ranking item i in position j. The average rank given each item is a popular summary which is a mean of these first order statistics.

Paul Holland suggested working with the exponential family through the first order statistics in the early 1970's. This leads to

Holland's model. Let ρ be the $n-1$ dimensional irreducible representation of S_n. Let $\mathrm{Mat}(n-1)$ be the set of all $n-1$ by $n-1$ real matrices. Define

$$P_\theta(\pi) = c(\theta) e^{Tr[\theta \rho(\pi)]}; \text{ for } \theta \in \mathrm{Mat}(n-1),$$

with $c(\theta)^{-1} = \Sigma_\pi \, e^{Tr(\theta(\rho(\pi))}$.

Remarks. These models are well parameterized by $\theta \in \mathrm{Mat}(n-1) = \mathbb{R}^{(n-1)^2}$. To give an example, consider a simple sub family:

$$Q_\theta(\pi) = c(\theta) e^{\theta \delta_1(\pi(1))}, \ \theta \in \mathbb{R}.$$

This can be described intuitively as "there is some special chance of ranking item 1 in position 1; whether or not this is done, the rest of the permutation is chosen uniformly.

If item 1 were carefully ranked, and then the others chosen at random, the appropriate family would be

$$Q_\theta(\pi) = c(\underset{\sim}{\theta}) e^{\theta_1 \delta_1(\pi(1)) + \ldots + \theta_{n-1} \delta_{n-1}(\pi(1))}, \ \underset{\sim}{\theta} \in \mathbb{R}^{n-1}.$$

Holland's model extends these considerations to a full first order model.

Joe Verducci (1982) began with Holland's model and the observation that $(n-1)^2$ parameters is still a lot to work with and think about. He introduced some natural low dimensional subfamilies and fit them successfully to real data sets. One of his nice observations is that some of Mallows' metric models introduced in Chapter 6-A-1 are subfamilies of first order exponential families.

Consider

$$Q_\lambda(\pi) = c(\lambda)e^{\lambda H(\pi,\pi_0)}\lambda \in \mathbb{R},\ H\ =\ \text{Hamming distance.}$$

For fixed π_0, this is a subfamily of Holland's, taking $\theta = \lambda\rho(\pi_0^{-1})$. Of course, if π_0 is also treated as a parameter, the two models are different. Verducci observed that replacing H by Spearman's S^2 also gives a first order model.

Arthur Silverberg (1980) began to work with second order models using the proportion ranking i, i' in position j, j'. Verducci (1982) realized the connection with group representations could help sort out questions of when a model is full, or well parameterized.

Silverberg worked with q-permutations, where people rank their favorite q out of n. This would be data on S_n/S_{n-q} in the language of Chapter 7. Generalizing slightly, let λ be a partition of n. Let $X = S_n/S_{\lambda_1} \times S_{\lambda_2} \ldots \times S_{\lambda_k}$ be the set or partial rankings of shape λ. Using Young's rule, and notation of Chapter 7,

$$L(X) = M^\lambda = \bigoplus_{\nu \vdash n} k(\nu;\lambda)S^\nu$$

where the sum is over all partitions ν of n which are larger than λ in the partial order of majorization and $k(\nu,\lambda)$ is the multiplicity of S^ν in M^λ. See the remarks to Theorem 1 in Chapter 7A. Restricting attention to a few of the pieces in this decomposition gives models of various sorts.

If $\lambda = (\lambda_1, \ldots, \lambda_k)$, the $n-1$ dimensional representation appears $(k-1)$ times $(k(n-1,1);\lambda) = k-1)$. The direct sum of these $k-1$ dimensional subspaces has dimension $(k-1)(n-1)$ and it spans the first order model.

Let us apply Young's rule to answer a question posed by Silverberg (1980) – what is the dimension of 2nd order models for q-permutation data. The partition involved is $n-q, 1^q$. Suppose that $2 \le q \le n-q$. Second order models are associated with partitions $(n-2,1,1)$ and $(n-2,2)$. By Young's rule, the multiplicity of each in $M^{1^q,n-q}$ is $\binom{q}{2}$. By the hook length formula of Chapter 7, the dimension of $S^{n-2,1,1}$ is $(n-1)(n-2)/2$. The dimension of $S^{n-2,2}$ is $n(n-3)/2$.

If we also include the first order component, the dimension of the second order model is

$$q\,(n-1) + \binom{q}{2}\binom{n-1}{2} + \binom{q}{2}\frac{n(n-3)}{2}.$$

Of course, it is important to keep the pieces separated, both for computation and inference.

The models discussed above have not been broadly applied. At present, there are no simple processes that lead to these models, nor simple interpretations or benefits from them. Since exponential families have such a good track record in

these directions, it seems like a worthwhile project to study and develop properties of low order exponential families on partially ranked data.

Some technical and practical aspects of the models in this section are discussed in Section E of this chapter.

D. Other models for ranked data.

The models proposed for ranked data in the previous section and the metric models of Chapter 6 have a distinctly ad-hoc flavor to them. There have been energetic attempts in the psychological literature to develop models for ranked data that are grounded in some more basic processes. This section briefly describes some of the models and gives pointers to the literature.

To fix a problem, consider an experiment in which p tones are played for a subject who is to rank them in order of loudness. It is an empirical fact that even a single subject, asked to repeat this task on different days, gives different answers. To account for this variability, Thurstone introduced an unobservable "discriminal process" of the form $u_1+X_1,\ u_2+X_2, \ldots u_p+X_p$ where $u_1,\ u_2, \ldots, u_p$ are fixed constants, and $X_1, \ldots, X_p$ are random variables, independent with the same distribution. It is postulated that on a given trial, a subject rank orders tone i in position j if $u_i + X_i$ is the jth largest.

Thurstone proposed normal distributions for the X_i. With a distribution fixed, one can estimate best fitting u_i and compare data and model. There has been a lot of experimental work showing a good fit for certain tasks. An extensive, readable review of this work appears in Luce and Suppes (1965).

A second line of work stems from a simple model put forward by Luce (1959). This postulates an unobservable system of weights $w_1, w_2, \ldots, w_p$. It is proposed that a subject ranks items by choosing the first ranked item with probability proportional to W_i. This choice being I, the second ranked item is chosen with probability proportional to $\{w_j\} - w_I$, and so on.

This model has also been fit to data with some success. Holman and Marley proved that if the underlying random variables X_i in Thurstone's approach have an extreme value distribution $P\{X < t\} = e^{-e^{-t}},\ -\infty < t < \infty$, the resulting choice probabilities are given by Luce model as well. Yellott (1977) gives references, proves a converse, and suggests some intriguing open probability problems.

Yellott's results deal with location shifts of extreme value distributions. Louis Gordon (1983) has observed a neat reformulation: consider the basic weights $w_1, \ldots w_p$ in Luce's model. Let $Y_1, Y_2, \ldots Y_p$ be independent and identically distributed standard exponential variables: $P(Y > t) = e^{-t}$. Put a probability on permutations by considering the order statistics of $Y_1/w_1, \ldots, Y_p/w_p$. Gordon shows this induces the distribution of Luce's sequential model. Since the log of an exponential variable has an extreme value distribution, this is a special case of the results described by Yellott. Gordon shows how to use the representation to give an efficient algorithm for generating random permutations from this distribution.

Independent of the literature cited above, Plackett (1975) developed a family of non-uniform probabilities on permutations. Plackett's first order models are the same as the Luce models. These are fit to some race horse data by Henery

(1981). An order statistics version of Plackett's higher order model is given by Dansie (1983). Plackett's motivation is interesting. One has available data on the chance that a horse finishes first in a race. One wants to predict the chance that the horse "shows" (finishes in the top 3). Plackett fit a model on the final permutation using the first order data. This approach is the basis of several *believable* systems for beating the races. See Zambia and Hausch (1984).

Models like Luce's have been extended, axiomatized, and tested by modern mathematical psychologists. The extensions account for practical difficulties such as the irrelevance of alternatives. If Luce's model is taken literally, one postulates a weight associated to the ith object independent of the other choices available. This easily leads to thought experiments generating data at variance with such a model. The following example is due to L. J. Savage.

Suppose you are indifferent between a trip to Paris and a trip to Rome. Thus w(Paris) $\doteq$ w(Rome). You clearly prefer Paris + \$10 to Paris. On Luce's model, if asked to choose between Paris, Paris + \$10, or Rome, you choose Rome about 1/3 of the time. Something is wrong here – it is unlikely that such a small inducement would change things so drastically. Tversky (1972) gives other examples and discussion.

One simple way around this objection is to allow the weights to depend on the problem under consideration. Going further, after the first choice is made, the second choice can be modeled by a new set of weights. But then any set of choice probabilities can be matched exactly so no test of the model is possible.

Some interesting half-way houses have been worked out. For example, Tversky (1972) describes choice by a hierarchical elimination process. Each alternative is viewed as a collection of measurable aspects. To make a choice, one selects an aspect with probability proportional to its measure. This eliminates all alternatives not possessing this aspect. The process continues until one alternative remains. For example, in choosing a restaurant for dinner, we may first choose type of food (e.g. seafood), then location, then price. Tversky and Sattath (1979) consider a subclass of these hierarchical models called preference trees which have many appealing properties.

The present state of the theory is this – no one claims to have a reasonable, believable and testable theory of how we perform ranking or choice. There *is* a list of constraints and desiderata on potential theories. These offer insight into choice behavior and rule out many naive suggestions. Thurstone's models and Luce's model are seen as straw men which triggered these investigations. Slight elaborations of these models have proven useful in horse race betting.

E. Theory and practical details.

1. *Justifying exponential families.*

Return to the setting of Section A – exponential families on a space X. One justification for these models P_θ that statisticians have developed goes as follows. Consider first a sample $X_1, X_2, \ldots X_n$ from such a family with unknown θ. The

sufficient statistics are

$$\bar{b}_i = \frac{1}{n}\sum_{j=1}^{n} b_i(X_j).$$

Any question about which $\theta \in \Theta$ generated the data can be answered as well from the averages $\bar{b}_i$ as from the full set of data. Often a working scientist, or common sense, will have reduced the data in just this way.

For example, if the data are n rankings of p items, it is natural to summarize the data by collecting together the number of people ranking item i in position j. This amounts to the first order models for permutations described in Section C above.

If summarization is deemed sensible, one may ask for the richest or fullest model for which this summarization is "legal." A classical theorem, the Koopman-Pitman-Darmois theorem, implies that this is the exponential family P_θ through these sufficient statistics.

This line of thinking has several modern versions. The Danish school of Martin-Löf-Lauritzen formalizes things as extreme point models. Lauritzen (1984) contains a clear description.

A Bayesian version is given by Diaconis and Freedman (1984). Briefly, if $X_1, X_2, \ldots, X_n$ (the data) are judged invariant under permutations (exchangeable) and more data of the same type could be collected, then de Finetti's theorem implies that the data were generated by a mixture of independent and identically distributed variables. *If* the $\bar{b}_i$ summarize the data, in the sense that given $\{\bar{b}_i\}$ all sequences $X_1, \ldots, X_n$ with these b_i are judged equally likely, then an extension of de Finetti's theorem implies the data are generated by a mixture of the exponential families introduced above. This brief description omits some technical details but is correct for the examples introduced below. Diaconis and Freedman also given versions of the Koopman-Pitman-Darmois theorem suitable for discrete data. Diaconis and Freedman (1988) give versions for continuous data.

There is a related motivation in the non Bayesian setting when x_i are iid: the maximum entropy distribution for $X_1, \ldots, X_n$ given the summaries $\{\bar{b}_i\}$ is the member $P_{\hat{\theta}}$ of the exponential family with $\hat{\theta}$ chosen so the mean of $P_{\hat{\theta}}$ equals $\bar{b}_i$. See Kullback (1968) or Posner (1975) for details.

These justifications boil down to the following: if the data are collected and it is judged reasonable to summarize by averages $\{\bar{b}_i\}$ then the exponential family P_θ gives the only probability model justifying this summary.

2. *Properties of exponential families.* Consider a sample $X_1, X_2, \ldots, X_n$ from P_θ, where it is assumed $\theta \in \mathbb{R}^p$. The maximum likelihood estimate of θ is a value $\hat{\theta}$ which maximizes $\Pi\, P_\theta(x_i)$. If X is finite this is an intuitively plausible procedure. It also has the Bayesian justification of being the (approximate) mode of the posterior distribution. Finally, it has quite a good track record in applied problems. The log-likelihood function is

$$L_n(\theta) = \theta' \sum_{i=1}^{n} b(X_i) - n\ \psi(\theta), \psi(\theta) = -\log\, a(\theta).$$

From the standard theory of maximum likelihood estimation in regular exponential families (see for example, Barndorf-Nielsen (1978) or Brown (1987)), we have

(i) $L_n(\theta)$ is strictly concave in θ.

(ii) $\psi(\theta)$ is analytic and $\bigtriangledown\ \psi(\theta) = E_\theta(b(x))$, $\bigtriangledown^2\ \psi(\theta) = \text{cov}_\theta(b(x))$, $\bigtriangledown^2\ \psi(\theta)$ is positive definite.

(iii) With probability one, there is an integer $n_0 = n_0(X_1, X_2, \ldots)$ such that the MLE $\hat{\theta}$ exists for all $n \geq n_0$. If the MLE exists, it is unique.

Crain (1974, 1976) gives results proving that, for continuous carriers,

- If the number of observations is larger than dim Θ, then the MLE exists.
- If dim Θ is allowed to grow with the sample size, then the "nonparametric density estimator" $f^*(x) = a(\theta^*)e^{\theta^* (x)}$ (θ^* the MLE) converges to the true sampling density. When X is finite this is clear, for eventually Θ becomes the set of all functions and $f^*(x)$ is then the frequency cell count for a multinomial.

(iv) A necessary and sufficient condition for the existence of the MLE is that $\bar{b}_i = \frac{1}{n}\sum_{i=1}^{n} b(X_i) \in$ int Hull (K), where K=range $\{b(x);\ x \in X\} \subset \mathbb{R}^p$.

(v) The MLE $\hat{\theta}$ exists iff the equations

$$E_\theta(b(X)) = \frac{1}{n}\sum_{i=1}^{n} b(X_i)$$

have a solution. When a solution exists it is unique and is the MLE. Thus, the MLE is that value of θ that makes the theoretical expectation of t equal its observed average.

(vi) The MLE is almost surely a consistent estimate of θ, and as n tends to infinity. Further, for large n, the difference between $\hat{\theta}$ and θ has an approximate normal distribution:

$$n^{\frac{1}{2}}(\hat{\theta} - \theta) \sim n(0,\ \bigtriangledown^2\psi(\theta)^{-1}).$$

This allows confidence intervals for θ, by using $\bigtriangledown^2\psi(\hat{\theta})^{-1}$ for the covariance matrix.

(vii) We have $P_\theta(dx) = a(\theta)e^{\theta' b}dx$. The sufficient statistics are $\bar{b}_i$. Following Crain (1974), consider a second expansion:

$$\frac{P_\theta(dx)}{dx} = \lambda_0 + \Sigma\ \lambda_i\ b_i(x).$$

If the b_i are orthogonal with respect to dx, then

$$\lambda_i = E_\theta(b_i) = E_\theta(\bar{b}_i).$$

In practice, $\hat{\theta}$ will not have a nice closed form expression. It will have to be determined numerically. There is a reasonable discussion of Newton-Raphson (called the method of scoring) in C. R. Rao's (1965) book. Beran (1979) suggests some other procedures as does Crain (1976).

There has not been a lot of work on a reasonable Bayesian analysis for these models. Consonni and Dawid (1985) develop some ideas which may generalize. A second starting place is to consider, as in Diaconis and Ylvisaker (1979), conjugate priors, and then their mixtures. There is probably some nice mathematics along the lines of Diaconis and Ylvisaker (1983), but bringing in some group theory.

3. *Introducing covariates.* A. P. Dempster (1971) has suggested a reasonable method of enlarging standard exponential families to include covariates. Suppose X is a finite homogeneous space. We observe pairs (x_i, z_i), $1 \leq i \leq n$ where $x_i \in X$ and $z_i \in \mathbb{R}^p$ is a covariate. Suppose that $b_1, b_2, \ldots, b_q$ is a basis for the model as above. The analog of Dempster's suggestion is the following family of probability densities (with respect to the uniform measure dx):

$$f(x|z) = \exp\left(\alpha + \sum_{i=1}^{q}\sum_{i=1}^{p} \phi_{ij} z_j b_i(x)\right).$$

Here, of course α is a normalizing constant and ϕ_{ij} are $p \cdot q$ parameters to be estimated. This amounts to the usual log-linear expansion

$$\exp\left(\alpha + \sum_{i=1}^{q} \alpha_i b_i(x)\right)$$

with $\alpha = \sum_{j=1}^{p} \phi_{ij} x_j$. Dempster discusses some of the calculus of such families, as well as some of the numerical and philosophical problems associated to such models. Dempster's analysis is an early version of the currently popular generalized linear model (GLM). See McCullagh and Nelder (1983). It may be that some of these analyses can be easily run in GLM.